Detailed Timeline of Main Events

Here is a detailed timeline of the main events and a cast of characters based on the provided sources:

Before 2010:

Development and initial use of traditional relational databases for structured data like transactional records and customer information.

Development of word embedding techniques like Word2Vec and GloVe for capturing semantic meaning in text data.

2010:

Wei Jin, A. O. publish a paper on using BERT for behavioral regression testing. This highlights early applications of advanced language models in specific technical domains.

2019:

Jacob Devlin, M.-W. C., et al. publish their paper on BERT (Bidirectional Encoder Representations from Transformers). This marks a significant advancement in language representation models with its ability to understand context from both directions in text.

2021:

Research explores techniques to scale Word2Vec on GPU clusters for faster processing (Jiaoyan mentioned in the text).

2022:

A study demonstrates Word2Vec's ability to identify meaningful relationships in ecological music and differentiate chord relationships (Izzidien mentioned in the text).

Word2Vec is used to analyze molecular cavities through topological vectorization.

Harsh Trivedi, N. B., et al. propose IRCoT, an approach interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. This is a precursor to more advanced RAG techniques.

July 2023:

AWS publishes a blog post discussing the role of vector databases in generative AI applications.

October 2024:

Databricks discusses Vector Databases in their glossary.

AWS Partner Network (APN) Blog highlights the importance of vector data stores for Gen AI Applications.

January 2025:

Several publications by **Satyadhar Joshi** discuss various aspects of GenAI, including financial applications, agent frameworks, workforce impact, and prompt engineering.

NVIDIA Blog publishes an article explaining Retrieval-Augmented Generation (RAG).

Jeffrey Pennington and R. S.'s work on GloVe is referenced (likely an ongoing resource).

Yang, B. L.'s work on text summarization based on BERT and GPT is referenced.

A report reviews data engineering and data lakes for implementing GenAI in financial risk (Satyadhar Joshi).

A paper reviews data pipelines and streaming for GenAI integration (Satyadhar Joshi).

A paper discusses implementing GenAI for increasing robustness of the US financial and regulatory system (Satyadhar Joshi).

A paper explores leveraging prompt engineering to enhance financial market integrity and risk management (Satyadhar Joshi).

A paper reviews data engineering frameworks (Trino and Kubernetes) for implementing Generative AI in Financial Risk (Satyadhar Joshi).

February 2025:

The IARJSET journal publishes "Introduction to Vector Databases for Generative AI: Applications, Performance, Future Projections, and Cost Considerations" by **Satyadhar Joshi**, providing a comprehensive overview of vector databases in the context of GenAI.

A paper reviews autonomous systems and collaborative AI agent frameworks (Satyadhar Joshi).

A paper reviews data engineering frameworks (Trino and Kubernetes) for implementing Generative AI in Financial Risk (Satyadhar Joshi).

A paper discusses advancing innovation in financial stability through AI agent frameworks (Satyadhar Joshi).

A paper analyzes agentic Generative AI and the future US workforce (Satyadhar Joshi).

A paper discusses mitigating workforce and economic disruptions with Generative AI (Satyadhar Joshi).

"Vector Database" is discussed under Graph Database & Analytics.

Access dates for online resources related to vector databases, Pinecone, and other technologies are mentioned, indicating ongoing relevance and information access.

2024 (Specific Month Not Always Specified):

Publication of "Artificial Intelligence Text Processing Using Retrieval-Augmented Generation: Applications in Business and Education Fields" by **Bogdan – Stefan POSEDARU, Florin – Valeriu PANTELIMON, M. – N. Dulgheru, and T. – M. Georgescu**, which discusses vector embeddings, word embedding models (Word2Vec, GloVe, BERT), and Retrieval Augmented Generation (RAG).

Publication of "Developing Concepts of Operations Using Multi-Step Tool Techniques with Large Language Models" by **Braxton VanGundy, Mikhail Schneide, Nipa Phojanamongkolkij, Ian Levitt, and Barclay Brown**, showcasing the use of LLMs with various tools, including databases, and mentioning several LLM models.

Bonan Min, H. R. publishes a survey on recent advances in NLP via large pre-trained language models.

Ongoing:

Development and increasing popularity of vector databases as critical infrastructure for AI applications, particularly generative AI.

Cloud providers like Google (Vertex AI, AlloyDB AI), AWS (vector data stores, Aurora), and Microsoft Azure (Cosmos DB, Semantic Kernel) invest heavily in and offer various vector database solutions.

The emergence of specialized vector databases (Pinecone, Qdrant, Milvus) and their integration with existing database systems and AI frameworks (LangChain, Vertex AI, OpenAI).

Discussions around the benefits and challenges of using vector databases, including scalability, performance, cost considerations, and the choice between specialized and integrated solutions.

The use of Retrieval-Augmented Generation (RAG) as a key technique to enhance the accuracy and relevance of Large Language Model responses by leveraging external knowledge stored in vector databases.
The Chartered Data Scientist (CDS™) credential program by ADaSci continues to offer certifications in data science and generative AI.
Pinecone, as a fully managed vector database, emphasizes ease of use, scalability, and low latency for AI applications.
The concept of semantic search, enabled by vector databases, becomes a foundational architecture for intelligent AI applications.

Cast of Characters and Brief Bios

Zachary Prer (Zach): Developer Advocate at Pinecone. His presentation focuses on explaining vector databases to application developers, highlighting their utility in solving real-world problems and enabling intelligent AI applications, particularly through semantic search and Retrieval-Augmented Generation (RAG). He emphasizes Pinecone's features like ease of use, scalability, and low latency.
Satyadhar Joshi: Independent Researcher (affiliated with BoFA, Jersey City, NJ, USA). Author of "Introduction to Vector Databases for Generative AI" and numerous other papers in 2025 focusing on the application of GenAI in finance, agent frameworks, workforce impact, and the role of vector databases. His work provides a comprehensive overview of the field.
Bogdan – Stefan POSEDARU: Researcher at Bucharest University of Economic Studies, Romania. Co-author of a paper on AI text processing using Retrieval-Augmented Generation, exploring the use of vector embeddings and RAG in business and education.
Florin – Valeriu PANTELIMON: Researcher at Bucharest University of Economic Studies, Romania. Co-author of the paper with Posedaru et al. on AI text processing and RAG.
M. – N. Dulgheru: Co-author of the paper on AI text processing and RAG.
T. – M. Georgescu: Co-author of the paper on AI text processing and RAG.
Braxton VanGundy: Researcher at NASA Langley Research Center. Lead author of a paper on developing concepts of operations using multi-step tool techniques with Large Language Models, showcasing practical applications of LLMs with various tools.
Mikhail Schneide: Researcher at NASA Langley Research Center. Co-author of the paper with VanGundy et al.
Nipa Phojanamongkolkij: Researcher at NASA Langley Research Center. Co-author of the paper with VanGundy et al.
Ian Levitt: Researcher at NASA Langley Research Center. Co-author of the paper with VanGundy et al.
Barclay Brown: Researcher at Collins Aerospace. Co-author of the paper with VanGundy et al.
Jeffrey Pennington: Researcher (affiliation mentioned as NLP Stanford Edu). Known for his work on GloVe (Global Vectors for Word Representation), a widely used word embedding technique.
Ryan Socher (R. S.): Associated with Jeffrey Pennington in the development of GloVe.
Jacob Devlin: Researcher. Lead author of the seminal paper on BERT (Bidirectional Encoder Representations from Transformers), a foundational model in NLP.
Ming-Wei Chang (M.-W. C.): Co-author of the BERT paper with Devlin et al.
Wei Jin: Author of a paper on using BERT for behavioral regression testing, highlighting an early application of BERT.

Alan Ouyang (A. O.): Co-author with Wei Jin on the paper about BERT for behavioral regression testing.
Bonan Min: Author of a survey paper on recent advances in Natural Language Processing via Large Pre-trained Language Models, highlighting the broader context in which vector databases and RAG are significant.
Harsh Trivedi: Author of a paper proposing IRCoT (Interleaving Retrieval with Chain-of-Thought Reasoning) for improved question answering, a concept related to RAG.
Niranjan Balasubramanian (N. B.): Co-author with Harsh Trivedi on the IRCoT paper.
Armen Izzidien: Researcher who demonstrated Word2Vec's ability to capture semantic relationships in music.
Jiaoyan: Researcher who explored scaling Word2Vec on GPU clusters for performance improvements.
Bolin Yang (B. L.): Author of work on recent progress in text summarization based on BERT and GPT.
ADaSci (Association of Data Scientists): A professional community offering memberships and accreditations like Chartered Data Scientist (CDS™) and Certified Generative AI Engineer, indicating a formal recognition and advancement within the data science and AI fields.
Pinecone: A company providing a fully managed, purpose-built vector database in the cloud, critical for enabling semantic search and Retrieval-Augmented Generation in AI applications.
Qdrant: An open-source vector database and vector search engine known for its flexibility in distance metrics and metadata filtering.
Weaviate: An open-source vector database that stores both objects and vectors, allowing for combined vector and structured search capabilities.
Milvus: An open-source vector database designed for scalable similarity search and high-performance with large datasets.
Google Cloud (Vertex AI, AlloyDB AI): A major cloud provider offering vector database services as part of its AI platform.
Amazon Web Services (AWS): A major cloud provider with various vector data store offerings.
Microsoft Azure (Cosmos DB, Semantic Kernel): A major cloud provider providing vector database capabilities within its cloud services and AI SDK.

This timeline and cast of characters provide a comprehensive overview of the main topics and individuals involved as presented in the provided sources.